IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Appellant:	Bin ZHANG	§	Confirmation No.:	5792
		§		
Serial No.:	10/694,367	§	Group Art Unit:	2168
		§		
Filed:	10/27/2003	§	Examiner:	J. A. Morrison
		§		
For:	Data Mining Method And System Using Regression Clustering	§ § §	Docket No.:	200310832-1

APPEAL BRIEF

Mail Stop Appeal Brief – Patents Date: April 9, 2007
Commissioner for Patents
PO Box 1450
Alexandria, VA 22313-1450

Sir:

Appellant hereby submits this Appeal Brief in connection with the above-identified application. A Notice of Appeal was filed via facsimile on February 16, 2007.

TABLE OF CONTENTS

I. REAL PARTY IN INTEREST

The real party in interest is the Hewlett-Packard Development Company (HPDC), a Texas Limited Partnership, having its principal place of business in Houston, Texas. The Assignment from the inventor to HPDC was recorded on October 27, 2003, at Reel/Frame 014652/0977.

II. RELATED APPEALS AND INTERFERENCES

Appellant is unaware of any related appeals or interferences.

III. STATUS OF THE CLAIMS

Originally filed claims:	1-30.
Claim cancellations:	None.
Added claims:	None.
Presently pending claims:	1-30.
Presently appealed claims:	1-30.

IV. STATUS OF THE AMENDMENTS

Appellant attempted to amend various claims after the Final Office Action dated October 17, 2006, but the Examiner did not enter the amendments.

V. SUMMARY OF THE CLAIMED SUBJECT MATTER

With the increase in the amount of data being stored in databases, the need to efficiently and accurately analyze data is increasing. Appellant's disclosure, para. [0002]. Appellant's contribution relates to techniques for efficiently mining data from datasets distributed across multiple locations.

According to the invention of claim 1, a processor-based method comprises selecting a set number of functions correlating variable parameters of a dataset. See e.g., Fig. 2, ref. no. 30 and para. [0025]. The method further comprises clustering the dataset by iteratively applying a regression algorithm and a K-Harmonic Means performance function on the set number of functions to determine a pattern in said dataset. See e.g., Fig. 2 and paras. [0025]-[0030].

According to the invention of claim 9, a storage medium comprises program instructions executable by a processor for selecting a set number of functions correlating variable parameters of a dataset. The program instructions also determine distances between datapoints of the dataset and values correlated with the set number of functions, calculate harmonic averages of the distances, regress the set number of functions using datapoint probability and weighting factors associated with the determined distances, repeating the determining and calculating for the regressed set of functions, compute a change in harmonic averages for the set number of functions prior to and subsequent to the regressing, and reiterating the regressing, repeating and computing upon determining the change in harmonic averages is greater than a predetermined value to thereby determine a pattern in the dataset. See e.g., Fig. 2 and paras. [0025]-[0030].

According to the invention of claim 15, a system comprises an input port configured to receive data and a processor configured to regress functions correlating variable parameters of a set of the data, cluster the functions using a K-Harmonic Mean performance function, and repeat the regressing and clustering sequentially to thereby determine a pattern in the dataset. See e.g., Fig. 2 and paras. [0025]-[0030].

According to the invention of claim 18, a system comprises a plurality of data sources and a means for regressively clustering datapoints from the plurality of data sources without transferring data between the plurality of data sources to thereby determine a pattern in data contained in said data sources. See e.g., Fig. 2 and paras. [0025]-[0030].

According to the invention of claim 24, a system comprises a plurality of data sources each having a processor configured to access datapoints within the respective data source and a central station coupled to the plurality of data sources and comprising a processor. The processors of the central station and plurality of data sources are collectively configured to mine the datapoints of the data sources as a whole without transferring all of the datapoints between the data sources and the central station to thereby determine a pattern in datapoints contained in the data sources. See e.g., Fig. 2 and paras. [0025]-[0030].

According to the invention of claim 28, a processor-based method for mining data comprises independently applying a regression clustering algorithm to a plurality of distributed datasets and developing matrices from probability and weighting factors computed from the regression clustering algorithm. The matrices individually represent the distributed datasets without including all datapoints within the datasets. The method further comprises determining global coefficient vectors from a composite of the matrices and multiplying functions correlating similar variable parameters of the distributed datasets by the global coefficient vectors to thereby determine a pattern in the datasets. See e.g., Fig. 2 and paras. [0025]-[0030].

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

Whether claims 1-30 are anticipated by Zhang et al. ("K-Harmonic Means-Data Clustering Algorithm," hereinafter the "Zhang Reference").

VII. ARGUMENT

A. Claims 1-17 and 28-30

Appellant selects claim 1 as representative of this claim grouping for purpose of the following argument. Claim 1 requires "iteratively applying a regression algorithm and a K-Harmonic Means performance function on the set number of functions to determine a pattern in said dataset." Claim 1 thus requires "regression." The Zhang Reference (authored by the present inventor) does not teach regression and, instead, refers to clustering. Clustering and regression are quite distinct. As proof of this point, Appellant submits the attached Table of Contents and Index from a well-known textbook entitled "Applied Regression Analysis" in the Evidence Appendix. Nowhere in the Table of Contents or in the Index of this regression-based textbook does a reference to "clustering" exist. This evidence proves that clustering and regression are distinct concepts.

Additionally, in the Zhang Reference, the clusters are represented by simple geometric centers. Each cluster is a subset of data surrounding a geometric point. In claim 1, however, clusters are represented by "functions" that correlate parameters of the dataset. As such, the claimed "functions" could be lines, curves, planes, hyperplanes, etc., not geometric centers.

Based on the foregoing, Appellant respectfully submits that the rejections of the claims in this first grouping be reversed, and the claims set for issue.

B. Claims 18-23

Appellant selects claim 18 as representative of this grouping. Claim 18 requires "regressively clustering datapoints." Because the Zhang Reference does not teach regression as explained above, the Examiner erred in rejecting claim 18. Based on the foregoing, Appellant respectfully submits that the rejections of the claims in this grouping be reversed, and the claims set for issue.

C. Claims 24-27

With regard to claim 24, the Examiner's Final Office Action quoted the claim language and simply pointed to page 1 of the Zhang reference. Independent claim 24 requires a plurality of data sources and a central station. Each of the plurality of data sources and the central station comprise a processor.

The claim further requires that the processors of the data sources and the central station "are collectively configured to mine the datapoints of the data sources as a whole without transferring all of the datapoints between the data sources and the central station." The Appellant has reviewed page 1 of the Zhang Reference, as well as the rest of the document, and simply does not find a teaching of this combination of limitations.

Based on the foregoing, Appellant respectfully submits that the rejections of the claims in this grouping be reversed, and the grouping set for issue.

D. Conclusion

For the reasons stated above, Appellant respectfully submits that the Examiner erred in rejecting all pending claims. It is believed that no extensions of time or fees are required, beyond those that may otherwise be provided for in documents accompanying this paper. However, in the event that additional extensions of time are necessary to allow consideration of this paper, such extensions are hereby petitioned under 37 C.F.R. § 1.136(a), and any fees required (including fees for net addition of claims) are hereby authorized to be charged to Hewlett-Packard Development Company's Deposit Account No. 08-2025.

Respectfully submitted,

Jonathan M. Harris
PTO Reg. No. 44,144
CONLEY ROSE, P.C.
(713) 238-8000 (Phone)
(713) 238-8008 (Fax)
ATTORNEY FOR APPELLANT

HEWLETT-PACKARD COMPANY
Intellectual Property Administration
Legal Dept., M/S 35
P.O. Box 272400
Fort Collins, CO 80527-2400

VIII. CLAIMS APPENDIX

1. (Previously presented) A processor-based method, comprising:

selecting a set number of functions correlating variable parameters of a dataset; and

clustering the dataset by iteratively applying a regression algorithm and a K-Harmonic Means performance function on the set number of functions to determine a pattern in said dataset.

2. (Original) The processor-based method of claim 1, wherein said clustering comprises:

determining distances between datapoints of the dataset and values correlated with the set number of functions;

regressing the set number of functions using datapoint probability and weighting factors associated with the determined distances;

calculating a difference of harmonic averages for the distances determined prior to and subsequent to said regressing; and

repeating said regressing, determining and calculating upon determining the difference of harmonic averages is greater than a predetermined value.

3. (Original) The processor-based method of claim 2, wherein said determining the distances comprises determining distances from each datapoint of the dataset to values within each function of the set number of functions.

4. (Original) The processor-based method of claim 2, wherein said selecting and said clustering are conducted for a plurality of datasets each from a different data source.

5. (Original) The processor-based method of claim 4, wherein said selecting and said clustering are conducted in parallel for each of the plurality of datasets.

6. (Original) The processor-based method of claim 4, further comprising determining a common coefficient vector to compensate for variations between similar sets of functions within the different data sources.

7. (Original) The processor-based method of claim 6, wherein said determining the common coefficient vector comprises:

developing matrices from the dataset datapoints and the probability and weighting factors for each of the datasets prior to said reiterating; and

determining the common coefficient vector from a composite of the developed matrices.

8. (Original) The processor-based method of claim 7, further comprising multiplying the similar sets of functions within the different data sources by the common coefficient vector.

9. (Previously presented) A storage medium comprising program instructions executable by a processor for:

selecting a set number of functions correlating variable parameters of a dataset;

determining distances between datapoints of the dataset and values correlated with the set number of functions;

calculating harmonic averages of the distances;

regressing the set number of functions using datapoint probability and weighting factors associated with the determined distances;

repeating said determining and calculating for the regressed set of functions;

computing a change in harmonic averages for the set number of functions prior to and subsequent to said regressing; and

reiterating said regressing, repeating and computing upon determining the change in harmonic averages is greater than a predetermined value to thereby determine a pattern in said dataset.

10. (Original) The storage medium of claim 9, wherein the program instructions are executable using a processor for computing the datapoint probability and weighting factors.

11. (Original) The storage medium of claim 9, wherein the program instructions are executable using a processor for developing matrices from the dataset datapoints and the probability and weighting factors prior to said reiterating.

12. (Original) The storage medium of claim 11, wherein the program instructions are executable using a processor for amassing matrices developed from a plurality of datasets each from a different data source.

13. (Original) The storage medium of claim 11, wherein the program instructions are executable using a processor for determining a common coefficient vector from the composite of matrices.

14. (Original) The method of claim 13, wherein the program instructions are executable using a processor for multiplying similar sets of functions within the different data sources by the common coefficient vector.

15. (Previously presented) A system, comprising:

an input port configured to receive data; and

a processor configured to:

regress functions correlating variable parameters of a set of the data;

cluster the functions using a K-Harmonic Mean performance function; and

repeat said regress and cluster sequentially to thereby determine a pattern in said set of data.

16. (Original) The system of claim 15, wherein the processor is arranged within one of a plurality of data sources each comprising a processor configured to:

regress the functions on a dataset of the respective data source;

cluster the functions using a K-Harmonic Mean performance function; and

repeat said regress and cluster sequentially.

17. (Original) The system of claim 15, further comprising a central station coupled to the plurality of data sources, wherein the central station comprises a processor configured to compute common coefficient vectors which compensate for variations between the regressively clustered functions representing the datasets, and wherein each of the processors of the data sources is configured to alter the functions by the common coefficient vectors.

18. (Previously presented) A system, comprising:

a plurality of data sources; and

a means for regressively clustering datapoints from the plurality of data sources without transferring data between the plurality of data sources to thereby determine a pattern in data contained in said data sources.

19. (Original) The system of claim 18, wherein the means for regressively clustering the datasets comprises a means for applying a regression algorithm and a K-Harmonic Means performance function on the datasets.

20. (Original) The system of claim 18, wherein the means for regressively clustering the datasets comprises a means for applying a regression algorithm and a K-Means performance function on the datasets.

21. (Original) The system of claim 18, wherein the means for regressively clustering the datasets comprises a means for applying a regression algorithm and an Expectation Maximization performance function on the datasets.

22. (Original) The system of claim 18, further comprising a central station communicably coupled to the plurality of data sources, wherein the means is further for:

collecting dataset information at the central station from the plurality of data sources;

determining a common coefficient vector from the collected dataset information; and

altering datasets within the plurality of data sources by the common coefficient vector.

23. (Original) The system of claim 18, wherein the means for regressively clustering the datasets comprises a storage medium with program instructions executable using a processor for:

selecting a set number of functions correlating variable parameters of a dataset;

determining distances between datapoints of the dataset and values correlated with the set number of functions;

regressing the set number of functions using datapoint probability and weighting factors associated with the determined distances;

calculating a difference of harmonic averages for the distances determined prior to and subsequent to said regressing; and

reiterating said regressing, determining and calculating upon determining the difference of harmonic averages is less than a predetermined value.

24. (Previously presented) A system, comprising:

a plurality of data sources each having a processor configured to access datapoints within the respective data source; and

a central station coupled to the plurality of data sources and comprising a processor, wherein the processors of the central station and plurality of data sources are collectively configured to mine the datapoints of the data sources as a whole without transferring all of the datapoints between the data sources and the central station to thereby determine a pattern in datapoints contained in said data sources.

25. (Original) The system of claim 24, wherein the each of the processors within the plurality of data sources is configured to regressively cluster a dataset within the respective data source.

26. (Original) The system of claim 25, wherein the processor within the central station is configured to:

collect information pertaining to the regressively clustered datasets; and

based upon the collected information, calculate common coefficient vectors which balance variations between functions correlating similar variable parameters of the regressively clustered datasets.

27. (Original) The system of claim 26, wherein the processor within the central station is further configured to:

compute a residual error from the common coefficient vectors;

propagate the common coefficient vectors to the data sources upon computing a residual error value greater than a predetermined value; and

send a message to the data sources to terminate the regression clustering of the datasets upon computing a residual error value less than a predetermined value.

28. (Previously presented) A processor-based method for mining data, comprising:

independently applying a regression clustering algorithm to a plurality of distributed datasets;

developing matrices from probability and weighting factors computed from the regression clustering algorithm, wherein the matrices individually represent the distributed datasets without including all datapoints within the datasets;

determining global coefficient vectors from a composite of the matrices; and

multiplying functions correlating similar variable parameters of the distributed datasets by the global coefficient vectors to thereby determine a pattern in said datasets.

29. (Original) The processor-based method of claim 28, further comprising repeating said independently applying, said developing, said determining and said multiplying.

30. (Original) The processor-based method of claim 28, further comprising calculating a residue error associated with the global coefficients prior to said multiplying.

IX. EVIDENCE APPENDIX

Applied Regression Analysis

THIRD EDITION

Norman R. Draper

Harry Smith

WILEY SERIES IN PROBABILITY AND STATISTICS
TEXTS AND REFERENCES SECTION

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

This book is printed on acid-free paper.

Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Draper, Norman Richard.
Applied regression analysis / N.R. Draper, H. Smith. — 3rd ed.
p. cm. — (Wiley series in probability and statistics. Texts and references section)
"A Wiley-Interscience publication."
Includes bibliographical references (p. —) and index.
ISBN 0-471-17082-8 (acid-free paper)
1. Regression analysis. I. Smith, Harry, 1923- . II. Title.
III. Series.
QA278.2.D7 1998

Contents

Preface to the Third Edition

The second edition had 10 chapters; this edition has 26. On the whole (but not entirely) we have chosen to use smaller chapters, and so distinguish more between different types of material. The tabulation below shows the major relationships between second edition and third edition sections and chapters.

Material dropped consists mainly of second edition Sections 6.8 to 6.13 and 6.15, Sections 7.1 to 7.6, and Chapter 8. New to this edition are Chapters 16 on multicollinearity, 18 on generalized linear models, 19 on mixture ingredients, 20 and 21 on the geometry of least squares, 25 on robust regression, and 26 on resampling procedures. Small revisions have been made even in sections where the text is basically unchanged. Less prominence has been given to printouts, which nowadays can easily be generated due to the excellent software available, and to references and bibliography, which are now freely available (either in book or computer form) via the annual updates in *Current Index to Statistics*. References are mostly given in brief either in situ or close by, at the end of a section or chapter. Full references are in a bibliography but some references are also given in full in sections or within the text or in exercises, whenever this was felt to be the appropriate thing to do. There is no precise rule for doing this, merely the authors' predilection. Exercises have been grouped as seemed appropriate. They are intended as an expansion to the text and so most exercises have full or partial solutions; there are a very few exceptions. One hundred and one true/false questions have also been provided; all of these are in "true" form to prevent readers remembering erroneous material. Instructors can reword them to create "false" questions easily enough. Sections 24.5 and 24.6 have some duplication with work in Chapter 20, but we decided not to eliminate this because the sections contain some differences and have different emphases. Other smaller duplications occur; in general, we feel that duplication is a good feature, and so we do not avoid it.

Our viewpoint in putting this book together is that it is desirable for students of regression to work through the straight line fit case using a pocket calculator and then to proceed quickly to analyzing larger models on the computer. We are aware that

X. RELATED PROCEEDINGS APPENDIX

None.